DOCUMENT RESUME

ED 442 825 TM 031 245

AUTHOR Reckase, Mark D.

TITLE The ACT/NAGB Standard Setting Process: How "Modified" Does It Have To Be before It Is No Longer a Modified-Angoff Process?

PUB DATE 2000-04-00

NOTE 11p.; Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 24-28, 2000).

PUB TYPE Reports - Descriptive (141) -- Speeches/Meeting Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.

DESCRIPTORS Competence; *Cutting Scores; Elementary Secondary Education; Evaluation Methods; National Surveys; Standards

IDENTIFIERS *Angoff Methods; *Standard Setting

ABSTRACT

The kernel of the Angoff method of standard setting (W. Angoff, 1971) would seem to be the judgment of whether a minimally competent person could answer an item on a test correctly or not. So it would seem that any procedure that requires independent judgment of the correctness or incorrectness of a response to items for a minimally acceptable examinee would merit being labeled as an Angoff method. Standard setting methods that use the Angoff kernel can be quite different in practice, but they may be given the same label in the literature on standard setting. The kernel for the modified Angoff method is not well-defined, but for this paper, it is defined as judgments of the proportion of correct responses for minimally qualified examinees on each item on a test. A look at the American College Test/National Assessment Governing Board (ACT/NAGB) standard setting process for 1992 shows that the definition of the kernel has evolved into something that only contains the essence of the modified Angoff kernel in the first round of ratings. Once the holistic information about test and item performance is provided following round one, it can no longer be argued that panelists are performing independent item-by-item ratings. Labeling a standard setting process by the summary description of the kernel is no longer adequate. The ACT/NAGB method is a multi-round process with a modified Angoff kernel for the first round, holistic feedback, and normative information for later rounds, and direct specification of cutscores on the score scale in the final round. (Contains 12 references.) (SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 442 825

M. Reckase

1

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

The ACT/NAGB Standard Setting Process: How "Modified" Does it Have to Be Before it is No Longer a Modified-Angoff Process?

Mark D. Reckase

TM031245

BEST COPY AVAILABLE

This paper is prepared for the:
Annual Meeting of the American Educational Research Association in New Orleans
April 2000

BEST COPY AVAILABLE

The ACT/NAGB Standard Setting Process: How "Modified" Does it Have to Be before it Is No Longer a Modified-Angoff Process?[1]

Mark D. Reckase
Michigan State University

Many standard setting processes have been developed over the years and are many schemes for categorizing these processes. Jaeger (1989) categorizes standard setting processes as being either test centered or student centered. Reckase (2000) suggests that recent surge in creative activity in standard setting demands a more flexible categorization scheme and he proposes a continuum of task demands as an underlying organization structure for classifying standard setting procedures. Whichever of many possible approaches for categorizing a standard setting procedure is used, the critical component of the standard setting process must be defined so that users know what is being classified. In philosophical terms, a standard setting procedure has an essence and it is the essence of the procedure that is classified. For the purposes of this paper, the basic judgement task that results in a standard setting process will be called the "kernel" of the standard setting process and it is the kernel that is categorized.

The kernel of a process can be used in unadorned form when standard setting is performed or the kernel can be embedded within a complex of supporting activities. Examples of these types of activities include: (1) training to perform the judgement task; (2) training on the policy definitions of the standards; (3) feedback to help inform judges about their level of understanding of the process; and (4) presentation of additional information, such as normative data, that is designed to place the standard setting process in a wider context. An operational standard setting process usually consists of the kernel plus the supporting activities. To accurately communicate the process to others, developers of standard setting processes need to clearly specify the kernel and describe the supporting components for the process.

The Angoff Standard Setting Process

The focus of this paper is not on standard setting processes in general, but on the Angoff standard setting process and modifications to that process (Angoff, 1971). What is the Angoff standard setting process and how do we know when we are using it? This question is important because the achievement levels set by the National Assessment Governing Board (NAGB) were criticized because the judgements required by the "Angoff procedure" were characterized as representing "a nearly impossible cognitive task." (p. 166) (Pelligrino, Jones,

[1] Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, April 2000.

& Mitchell, 1999). If this criticism is taken seriously, it is important that practitioners know when an Angoff procedure is being used. So, what is the kernel of the Angoff procedure?

The original description of the Angoff procedures is given in one paragraph in Angoff (1971). That paragraph is reproduced here.

> "A systematic procedure for deciding on the minimum raw scores for passing and honors might be developed as follows: keeping the hypothetical 'minimally acceptable person' in mind, one could go through the test item by item and decide whether such a person could answer correctly each item under consideration. If a score of one is given for each item answered correctly by the hypothesized person and a score of zero is given for each item answered incorrectly by that person, the sum of the item scores will equal the raw score earned by the 'minimally acceptable person.' A similar procedure could be followed for the hypothetical 'lowest honors person.'"

The kernel of this procedure would seem to be the judgement of whether a minimally competent person could answer an item on a test correctly or not. So it would seem that any procedure that requires independent judgements of the correctness or incorrectness of a response to items for a minimally acceptable examinee would merit being labeled as an Angoff method.

Unfortunately, developers of new standard setting methods do not always thoroughly review the literature before presenting their new method. Reckase and Bay (1998) called a method with the Angoff kernel the "item score string method" (ISSE). Impara and Plake (1997) and Verhelst and Kaftandjiva (1999) called the Angoff kernel the yes/no method and the item mastery method, respectively. Yet, all of these methods require the same type of judgement, whether a person at a minimal level of proficiency could answer the items correctly.

The supporting activities for the Angoff kernel are seldom described at any level of detail. The result is that standard setting procedures that use the Angoff kernel can be quite different. One procedure might have one round of judgements with no feedback while another might have multiple rounds with feedback on location of the cutscore and an estimate of the proportion of examinees that will exceed the cutscore. These processes have the same kernel, but they are very different in practice. Yet, they will very likely be given the same label when referenced in the literature on standard setting.

The Modified Angoff Standard Setting Process

The kernel for the modified Angoff method for standard setting is far from well defined. Angoff (1971, p 515) suggested a variation to the procedure described above in a footnote on the same page.

> "A slight variation of this procedure is to ask each judge to state the *probability* that the 'minimally acceptable person' would answer each item correctly. In effect, the judges would think of a number of minimally acceptable persons, instead of only one such person, and would estimate the proportion of minimally acceptable persons who would answer each item correctly. The sum of these probabilities, or proportions, would then represent the minimally acceptable score."

Since the judgement of probability of a correct response for a minimally competent examinee is a variation on judging whether a correct or incorrect response would be given, this procedure seems to be a modification of the initial procedure. But Shepard (1984) labels the procedure based on estimating probabilities as *the* Angoff procedure and that based on judging correct or incorrect responses as the modified-Angoff approach.

Adding to the confusion, Jaeger (1989) describes two modifications to the Angoff procedure. One gives multiple choices of probabilities of correct responses for the judges to select from rather than making direct probability estimates. The second modification is adding a discussion of the specified probabilities of correct response between the two rounds of probability judgements. Jaeger (1989) considers estimating the proportion of minimally acceptable persons who would answer each item correctly as *the* Angoff procedure.

Guion (1998) continues the pattern of labeling judgements of the probability of correct response for minimally competent candidates as the Angoff process. But Guion describes a different process as the modified Angoff process. For him, the modified Angoff process involves adjusting the judgements using data on previous performance on the items. It is clear from the review of these few sources, that there is no consensus on the definition for the modified Angoff process, or even for the definition of the original Angoff process.

Rather than try to resolve what seems to be a pervasive level of confusion about the characteristics of the Angoff procedure, the kernel of the modified Angoff procedure will be defined here as judgements of the proportion of correct responses for minimally qualified examinees on each item on a test. The judgements of the proportion of correct responses are done independently for

each item on the test. This definition of the kernel of the modified Angoff method is consistent with the focus of the critical comments made in Pellegrino et al (1999). Procedures that use this type of item-by-item rating will be labeled as having a modified Angoff kernel. Those that stray into other types of judgements will be given a different label.

The classification of whether or not something is a modified Angoff process becomes difficult if an operational standard setting process has a complex supporting structure. For example, if judges get information about the proportion of examinees that will be above the cutscore, does this change the kernel of the process? Is the standard setting method now holistic because there may be a tendency to raise or lower the cutscore to get a desired passing proportion? This issue is addressed in the next section of this paper.

The 1992 ACT/NAGB Standard Setting Process

The first operational standard setting performed by ACT, Inc. for NAGB was done for NAEP Mathematics in March, 1992 (American College Testing, 1993). The question to be addressed here is whether the process as implemented can accurately be labeled as a modified Angoff process as defined in this paper. That is, was the standard setting accomplished using an item-by-item rating method that required participants to indicate the proportion of minimally qualified students who would answer each item correctly? To answer that question, the detailed description of the 1992 NAEP Mathematics standard setting process, now called Achievement Levels Setting, will be reviewed.

The 1992 NAEP Mathematics Achievement Levels Setting process began with thorough training on NAEP, NAGB, and the achievement levels setting process. Next participants, called panelists in descriptions of the process, took the NAEP test for the grade level they were considering. Following these training tasks, the panelists worked together to translate the policy statements about the standards that were to be set, called policy definitions, into content specific skill and knowledge descriptions called achievement level descriptions (ALDs). The ALDs were used to bring about a common understanding of the skills and knowledge possessed by minimally qualified at an achievement level. The ALDs guided the rest of the process.

Before doing any rating of items, the panelists were given thorough training on the process. The rating process consisted of the following:

> "For the multiple-choice and short-answer items, panelists were asked to make a judgement as to how many of those 100 students, at each borderline achievement level, would answer the item correctly. And for the extended response items, panelists were asked to review 20 to 25 actual student responses (papers) and select three papers, one for each achievement level, that typified student performance at the borderline of that level." (American College Testing, 1993, p. 22)

Panelists performed three rounds of ratings with supporting information provided between rounds. After the first round of ratings, panelists were given feedback about the location of the standard that they had set and those of the other panelists in their grade group. In addition, they were given information about the proportion of all students who took the items that answered them correctly for the items scored 0,1, and the distribution of item scores on the extended response items. After digesting this information, they reviewed the round one ratings and either decided to leave them as they were or adjusted them to take into account the information they had been given.

Following the round two ratings, panelists were again given information about the location of the standards they had set and those of the other panelists. They were also given lists of items that were rated differently than items of similar difficulty. They might have misinterpreted these items. They also had access to the information about item difficulty and the distribution of scores on the extended response items.

The rating of the items scored 0 or 1 in the 1992 NAEP Mathematics ALS process is clearly consistent with the definition of the Angoff process given above. The standard setting for the extended response items was based on the selection of actual papers from sets of 20 to 25 actual papers for each extended response item. While this is an item-by-item rating process, it does not have the same kernel as the modified Angoff method. However, there were relatively few extended-response items so it would seem that overall the 1992 NAEP Mathematics ALS process can be described as having a modified Angoff kernel. The process also had extensive supporting activities including a variety of different types of feedback and some information about item difficulty.

The 1998 ACT/NAGB Standard Setting Process

ACT has conducted a number of standard setting processes for NAGB since 1992. Each has been somewhat different than the previous one. To emphasize the evolution of the processes, the 1998 NAEP Civics Achievement Levels Setting process is described in comparison to the 1992 process. The 1998 process continued the extensive training about NAEP, NAGB, and the standard setting method. Rather than develop content descriptions during the process, they were developed by an earlier committee and were discussed during this process. A different process was used for setting standards with the extended-response items because there were many more of items of this type on civics than on 1992 NAEP Mathematics, but selecting student papers as examples of the minimally qualified students was still done as part of training.

The standard setting process was extended to four rounds. The additional round was included so that information about the proportion of students estimated to be above the cutscores could be provided without influencing the earlier process. The components of the rounds of rating and the information provided between rounds are described below.

The round one rating process consisted of estimating the proportion of students at the lower borderline of an achievement level that would answer a dichotomously scored items correctly, and the mean scores for the borderline students on items that had multiple score points. The estimation of mean scores on extended response items replaced the paper-selection method that had been used for mathematics.

Following the round one ratings, a variety of information was provided. The panelists received information about the location of their cutscores and the others in their grade group. They also received examples of test booklets for students who obtained scores near the cutscore that had been set. This information was provided so that panelists would have a better understanding of the standard that they had set. The panelists also received the proportion of students who answered each item correctly and the score distribution on the extended response items. Finally, they received a summary of their ratings on Reckase Charts, a means for showing their ratings relative to the characteristics of the items as determined from the IRT calibration (see Loomis (2000) for a more complete description of these types of feedback). Four-and-one-half hours were allocated to reviewing the information following round one.

Round two used the same rating process as round one. Panelists could leave their ratings as they were, or could change them to reflect changes in their understanding of student performance brought about by the information they were provided. Following round two, panelists received the same kinds of

information they had received previously, and information about the proportion of students who were estimated to fall above the cutscores set by their grade group.

Round three ratings used the same process as the earlier rounds. Following that round, panelists again received information about the proportion of students above the cutscores, but now the information was provided for the standards set by each panelist and for each grade level. The final round of rating was done directly on the reporting score scale. Panelists were given the opportunity to change the cutscores directly without going through the item by item rating process.

For the purposes of this paper, the presentation of this process has necessarily been brief. The entire 1998 NAEP Civics ALS process took five days and it is impossible to describe it in detail in a short paper. The details can be obtained from ACT reports of the process (e.g., ACT, 1997). While this description is somewhat abbreviated, it should still be possible to determine whether the 1998 NAEP Civics process is a modified Angoff process.

Is the Current ACT/NAGB Process a Modified Angoff Process?

The essence of the modified Angoff process is the estimation of the proportion of minimally qualified examinees that can answer each item correctly. The first three rounds of the 1998 NAEP Civics process appear to use that kernel for the items scored correct or incorrect. But there is also the mean estimation for the extended response items. Does that method have a different kernel than the modified Angoff kernel? The modified Angoff kernel seems to require independent ratings of each item. Do whole booklet feedback and Reckase Charts affect independence of panelists' ratings? Both types of information are designed to encourage panelists to look at the standards relative to performance on the test as a whole rather than item by item. Then there is the obvious deviation from the modified Angoff kernel that results from the direct specification of the standard on the score scale in round four of the procedure after receiving information about the proportion of students estimated to be above each cutscore.

My own evaluation of the ACT/NAGB process as currently implemented is that it has evolved into something that only contains the essence of the modified Angoff kernel in the first round of ratings. Once the holistic information about test and item performance is provided following round one, it can no longer be argued that panelists are performing independent item-by-item ratings. Certainly, panelists provide ratings for each item, but the focus shifts to adjusting cutscores on the score scale. The standards panelists set should result in booklets that represent the panelists' conception of minimally acceptable

performance. Also, the percent of students estimated to be above a cutscore provides normative information and the direct specification of cutscores defines an entirely different kernel for a standard setting process. Although I would like to continue to honor Angoff as the author of a practical and widely used standard setting procedure, the ACT/NAGB process has evolved far beyond the method described in two paragraphs in 1971.

A New Direction

The elaborate complexity of the ACT/NAGB process indicates that labeling a standard setting process by the summary description of the kernel is no longer adequate. Processes with the same kernel differ dramatically. At the very least, the description of a standard setting process should contain the following information.

1. The policy statements that call for the standard.
2. The process for selecting the panelists and the qualifications of the panelists.
3. The number of rounds of ratings and the kernel of the process used in each round.
4. A thorough description of the information provided to panelists between rounds of ratings.
5. The amount and types of training given the panelists to prepare them to perform the process.

The ACT/NAGB method is a multi-round process with a modified Angoff kernel for the first round, holistic feedback and normative information for later rounds, and direct specification of cutscores on the score scale in the final round. The description could include a summary of policy and the selection process for panelists as well, but that is beyond the scope of this paper. Calling the ACT/NAGB process a modified Angoff process is like calling a seven-course banquet a supper. There may be a small amount of truth in the labeling, but such a characterization does not capture the elaborate nature of the meal and the care with which it was prepared. The same is true for the ACT/NAGB process.

References

ACT (1997, December). *Developing achievement levels on the 1998 NAEP in civics and writing: Design document.* Iowa City, IA: Author.

American College Testing (1993, February). *Description of mathematics achievement levels-setting process and proposed achievement level descriptions: 1992 National Assessment of Educational Progress (Volume I).* Iowa City, IA: Author.

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.) *Educational measurement (2nd Edition).* Washington, DC: American Council on Education.

Guion, R. M. (1998). *Assessment, measurement, and prediction for personnel decisions.* Mahwah, NJ: Lawrence Erlbaum Associates.

Impara, J. C., & Plake, B. S. (1997). Standard setting: an alternative approach. *Journal of Educational Measurement, 34(4),* 353-366.

Jaeger, R. M. (1989). Certification of student competence. In R. L. Linn (Ed.) *Educational measurement (3rd Edition).* New York: American Council on Education and Macmillan.

Loomis, S. C. (2000, April) Feedback in the NAEP achievement levels setting process. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans.

Pellegrino, J. W., Jones, L. R., & Mitchell, K. J. (Eds.) Grading the nation's report card: Evaluating NAEP and transforming the assessment of educational progress. Washington, DC: National Academy Press.

Reckase, M. D. (2000). A survey and evaluation of recently developed procedures for setting standards on educational tests. Paper written on contract to the National Assessment Governing Board.

Reckase, M. D., & Bay, L. (1998, April). Analysis of methods for collecting test-based judgements. Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego.

Shepard, L. A. (1984). Setting performance standards. In R. K. Berk (Ed.) *A guide to criterion-referenced test construction.* Baltimore: Johns Hopkins University Press.

Verhelst, N. D., & Kaftandjieva, F. (1999). *A rational method to determine cutoff scores (Research Report 99-07*). Enschede, The Netherlands: University of Twente.